Evaluation of Standards data collected from probabilistic sampling programs

Eric P. Smith

Y. Duan, Z. Li, K. Ye

Statistics Dept., Virginia Tech

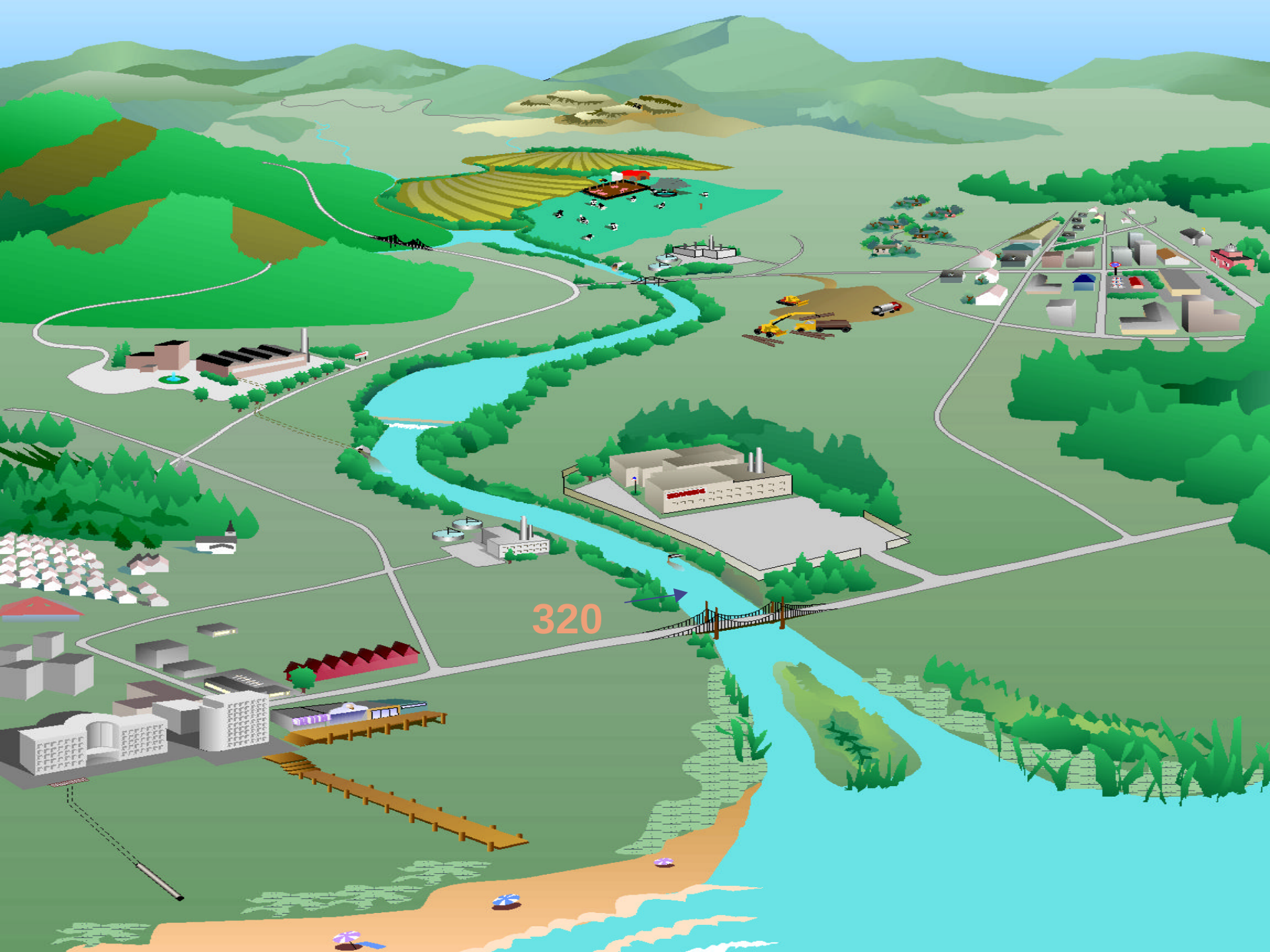
Sponsor

This research is funded by
U.S. EPA - Science To Achieve Results (STAR) Program
Grant # RD-83136801-0

This talk was not subjected to USEPA review. The conclusion and opinions are soley those of the authors and not the views of the Agency.

Outline

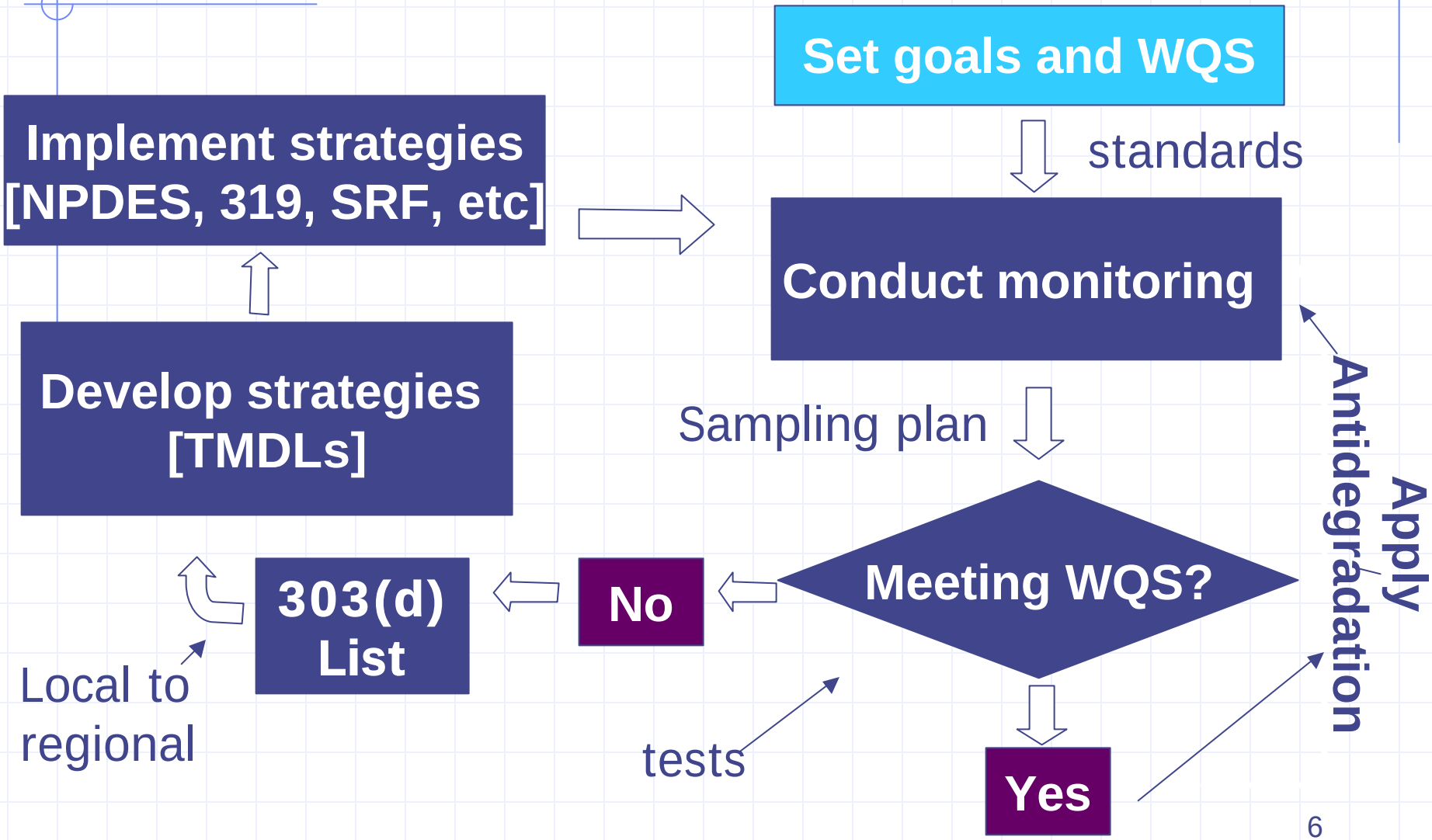
- Background
  - Standards assessments
- Single site analysis
- Regional analysis
  - Mixed model approach
  - Bayesian approach
- Upshot: need models that allow for additional information to be used in assessments


320

Standards assessment – 303d

- Clean Water Act section 303d mandates states in US to monitor and assess condition of streams
- Site impaired – list site, start TMDL process (Total Max Daily Loading)
- Impaired means site does not meet usability criteria

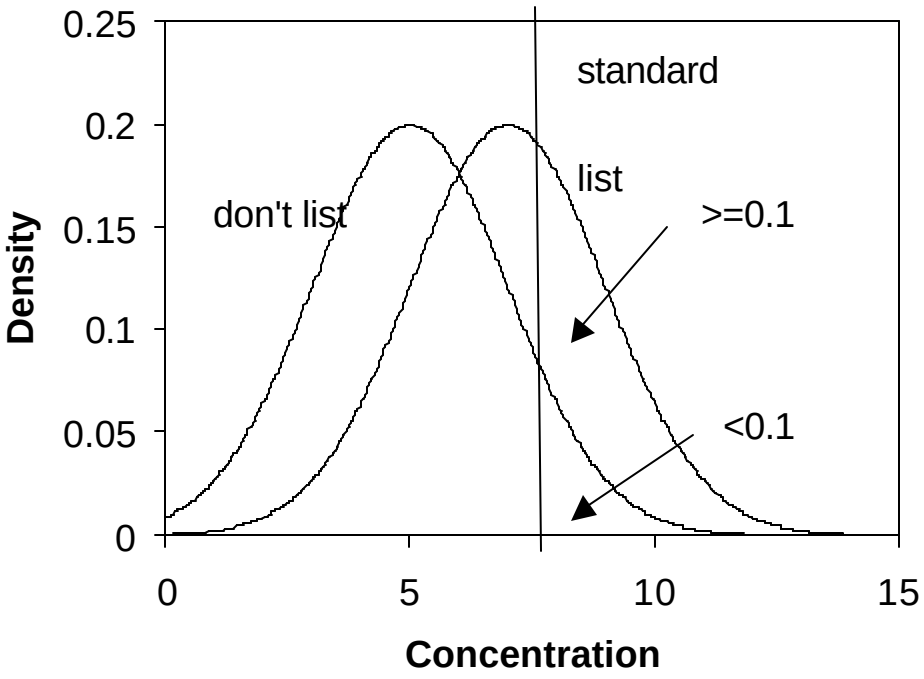
Linkages in 303(d)



Impaired sites

- Site impaired if standards not met
- Standards – defined through numerical criteria
  - Involve frequency, duration, magnitude
- –Old method
  - Site impaired if >10% of samples exceed criteria
  - Implicit statistical decision process- error rates

Test of impairment



Newer approach to evaluation

- Frequency:
  - Binomial method
  - Test $p<0.1$
- Magnitude
  - Acceptance sampling by variables
  - Tolerance interval on percentile
  - Test criteria by computing mean for the distribution of measurements and comparing with what is expected given the percentile criteria

Problems

- Approach is local
  - Limited sampling budget; many stations means small sample sizes per station
  - Impairment may occur over a region
  - Modeling must be relatively simple (hard to account for seasonality, temporal effects)
  - Does not complement current approaches to sampling
  - Site history is ignored
  - Not linked to TMDL analysis (regional) and 305 reporting

Probabilistic sampling schemes

- Rotating panel surveys
  - Some sites sampled at all possible times
  - Other sites sampled on rotational basis
  - Sites in second group may be randomly selected

Making the assessment regional

Y = mean + site
Y = mean + time + site

$$\boldsymbol{y} = X\boldsymbol{ß} + Z\boldsymbol{u} + \boldsymbol{e}$$

- X defines fixed effects (time), Z defines random ones (site, location), β, u are parameters
- Covariances

$$\boldsymbol{e} \sim MVN(0,\Gamma)$$
$$\boldsymbol{u} \sim MVN(0,G)$$

$$V(y) = ZGZ' + \Gamma$$

Regional Mixed Model

- Allows for covariates
- Allows for a variety of error structures
  - Temporal, spatial, both
- Does not require equal sample sizes etc
- Allows estimation of means for sites with small sample sizes
  - Improves estimation by borrowing information from other sites

Simple model

$$y_{ij} = m + a_i + e_{ij}$$

Error term allows for modeling of temporal or spatial correlation

Random site effect

- Testing is based on estimate and variance of mean for site i (μ_i)
- Can also test for regional impairment using distribution of grand mean

Error and stochastic components

$$y_{ij} = m + a_i + e_{ij}$$

Error term allows for modeling of temporal or spatial correlation

Random site effect

- Covariance Structure without correlation (one random effect model)

$$e_{ij} \overset{iid}{\sim} N(0, s^2)$$

- Spatial Covariance Structure

$$Var(e) = \sigma^2 \begin{bmatrix} 1 & \rho^{d_{12}} & \rho^{d_{13}} & \rho^{d_{14}} \\ \rho^{d_{21}} & 1 & \rho^{d_{23}} & \rho^{d_{24}} \\ \rho^{d_{31}} & \rho^{d_{32}} & 1 & \rho^{d_{34}} \\ \rho^{d_{41}} & \rho^{d_{42}} & \rho^{d_{43}} & 1 \end{bmatrix}$$

Test based on OLS estimations for each site i

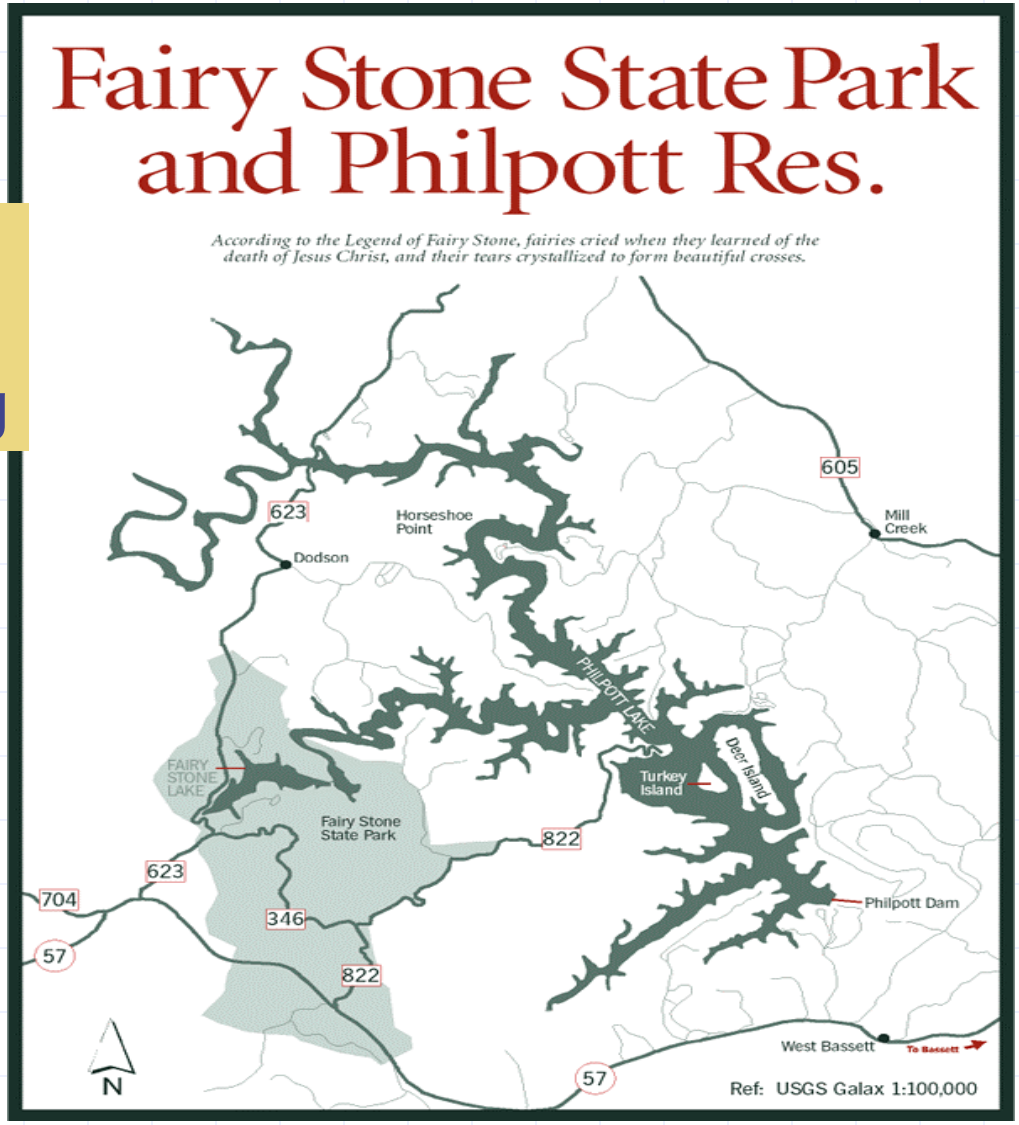
$$\frac{\bar{y}_i - baseline}{\hat{s} / \sqrt{n_i}} \sim t_{df,d}$$

where $\bar{y}_i$ and $\hat{s}$ are OLS estimates of m and s;

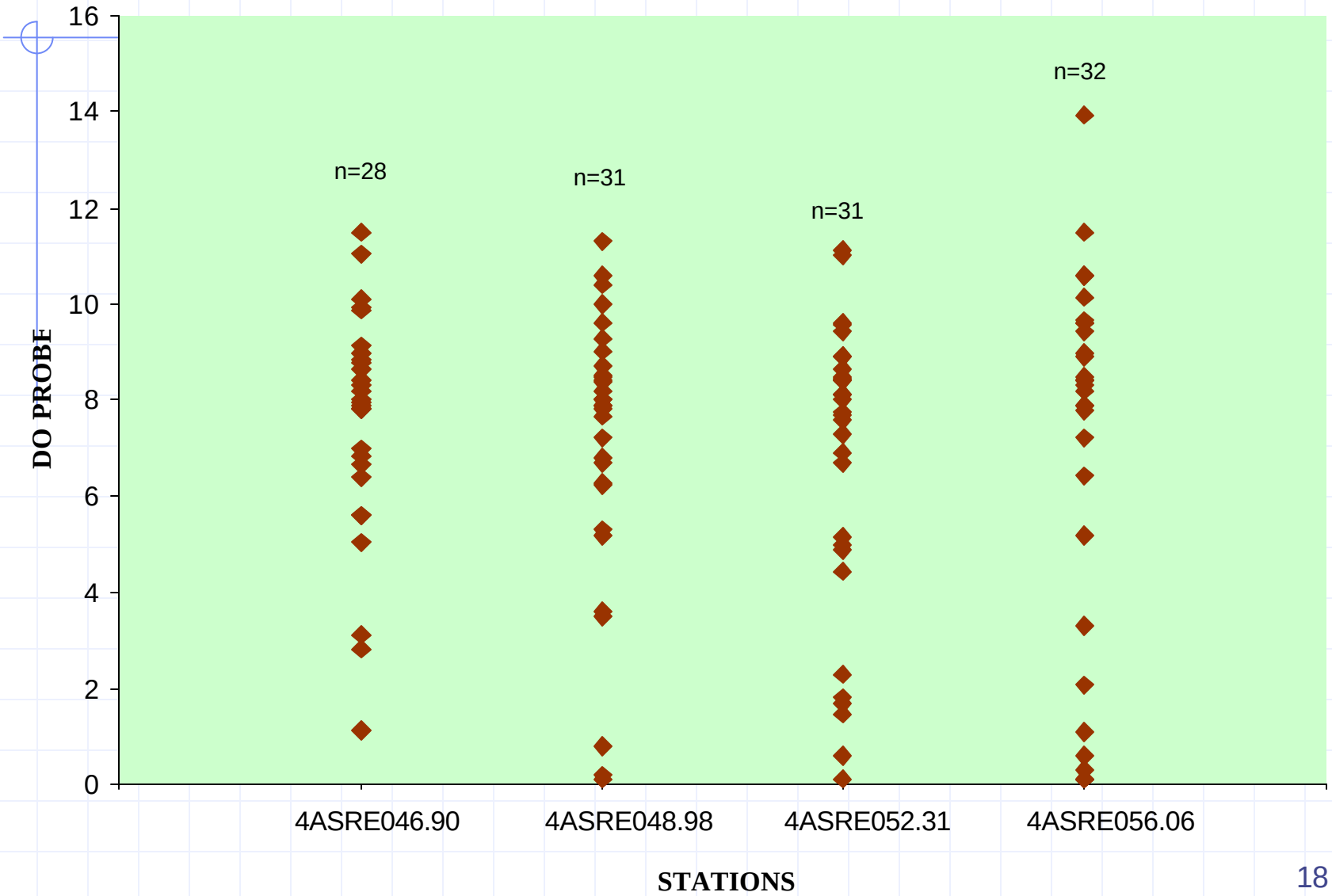
$df = n_i - 1,\ d = noncentrality$

- Baseline is the standard. For DO, we use 5, and for PH 6.
- Model based: same idea but mean and variance are estimated from model

Located in SW Virginia
Good bass fishing



DO data collected at four stations of PHILPOTT RESERVOIR (years 2000, 2001 & 2002)



Evaluation based on Do data of PHILPOTT RESERVIOR (2000-2002)

	4ASRE046.90	Model based	4ASRE052.31	4ASRE056.06
n	28		31	32
Sample mean	7.55		6.66	6.67
Sample variance	5.81		9.56	16.15
% excceding	11		26	28
Binomial p-value	.5406		.0096	.0033
Test statistic	5.6	4.27	2.99	2.35
critical value	4.75	5.05	5.19	5.2
conclusion	**Fail to reject**	**reject**	reject	reject

Single site analysis

Bayesian approach

$$y_{ij} = m + a_i + e_{ij}$$

- a is a random site effect
- Error term may include temporal correlation or spatial
- Priors on parameters
  - Mean –uniform
  - a is normal (random effect) variance has prior

$$p(s^2, s_a^2) \propto \frac{1}{s^2} \frac{1}{s^2 + s_a^2}$$

Alternative: Using historical data

- Power prior – Chen, Ibrahim, Shao 2000
- Use likelihood from the previous assessment (D_0). Basic idea: weight new data by prior data
- Power term, d , determines influence of historical data.
- Modification to work with Winbugs

Incorporate Historical Data using Power Priors

- Make d random, and assign a prior $p(d) = Beta(a,b)$ on it. The joint posterior of (q,d) becomes

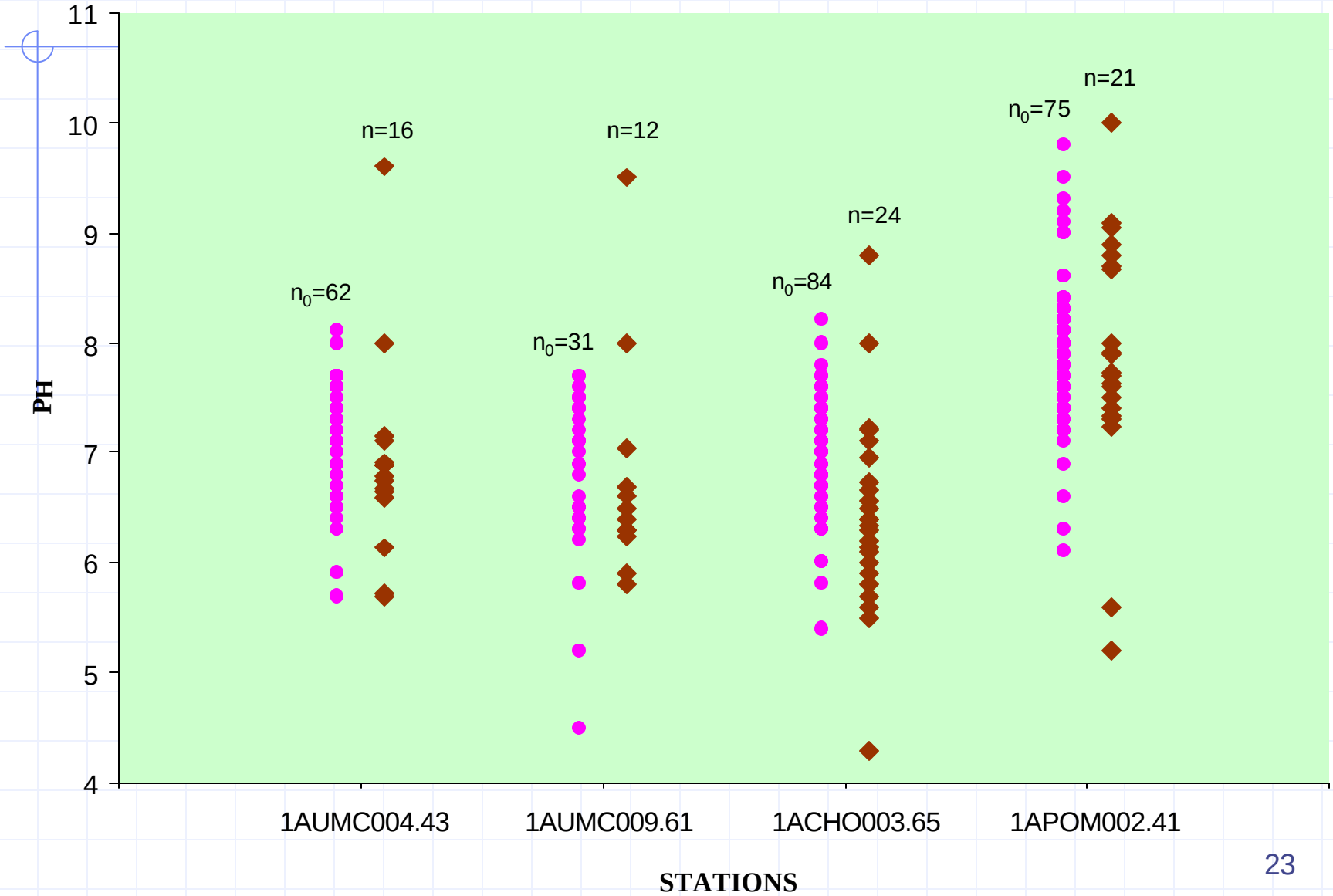
$$p(q,d \mid D_0, D) \propto \frac{L(q \mid D)(L(q \mid D_0))^d p(q) p(d)}{\int (L(q \mid D_0))^d p(q) dq} I_A(d)$$

where D is current data and D0 is past data

$$A = \left\{ d : 0 < \int p(q) \left(L(q \mid D_0) \right)^d dq < \infty \right\}$$

- Advantage: Improve the precision of estimates.

PH data collected at four stations: use past information to build prior



Evaluate site impairment based on PH data with power priors

Station of interest	1AUMC004.43	1AUMC009.61	1Acho003.65	1APOM002.41
n	16 (yr.99-02)	12 (yr.99-01)	24 (yr.99-01)	21 (yr.99-00)
No. obs <6	2	2	6	2
sample mean	6.91	6.78	6.43	7.87
sample variance	0.82	1.06	0.78	1.23
n_0	62 (yr.90-98)	31 (yr.90-98)	84 (yr.90-98)	75 (yr.90-98)
sample mean of D_0	7.05	6.73	6.95	7.88
Percent exceed the EPA standard	0.13	0.17	0.25	0.10
P-value of Binomial test (H_0: p=0.1 H_a: p>0.1)	0.4853	0.3410	0.0277	0.6353
Bayesian test. (H_0: L=6 H_a: L<6), L is the lower 10th percentile of PH				
With Reference Prior:				
$P(H_0)$	0.1663	**0.0502**	0.0003	0.8673
posterior s.d. of ?	0.3399	0.4708	0.262	0.3564
With Power Prior:				
$P(H_0)$	0.4868	**0.03525**	0.0017	0.9831
posterior s.d. of ?	0.2566	0.2562	0.2381	0.2477

Power Priors with Multiple Historical Data Sets

- If multiple historical data sets are available, assign a different d_j for each historical data set.

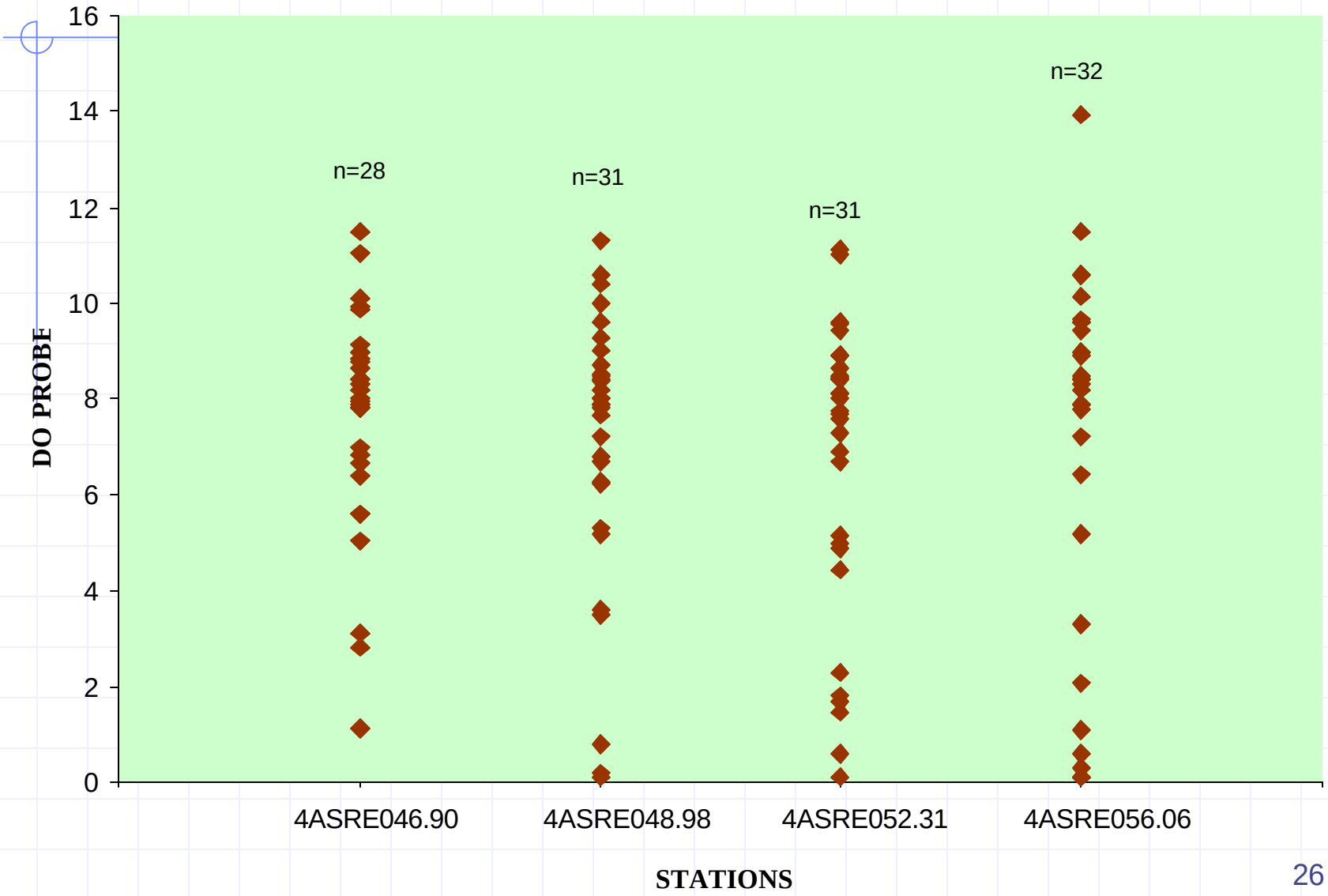
$$p(q,\underline{d} \mid D_0, D) \propto \frac{L(q \mid D)\left(\prod_{j=1}^{m}(L(q \mid D_{0j}))^{d_j} p(d_j)\right) p(q)}{\int\left(\prod_{j=1}^{m}(L(q \mid D_{0j}))^{d_j}\right) p(q)\,dq} I_B(\underline{d})$$

where

$$B = \left\{(d_1,\ldots,d_m) : 0 < \int\left(\prod_{j=1}^{m}\left(L(q \mid D_{0j})\right)^{d_j}\right) p(q)\,dq < \infty\right\}$$

- Data collected at adjacent stations could be used as "historical" data.

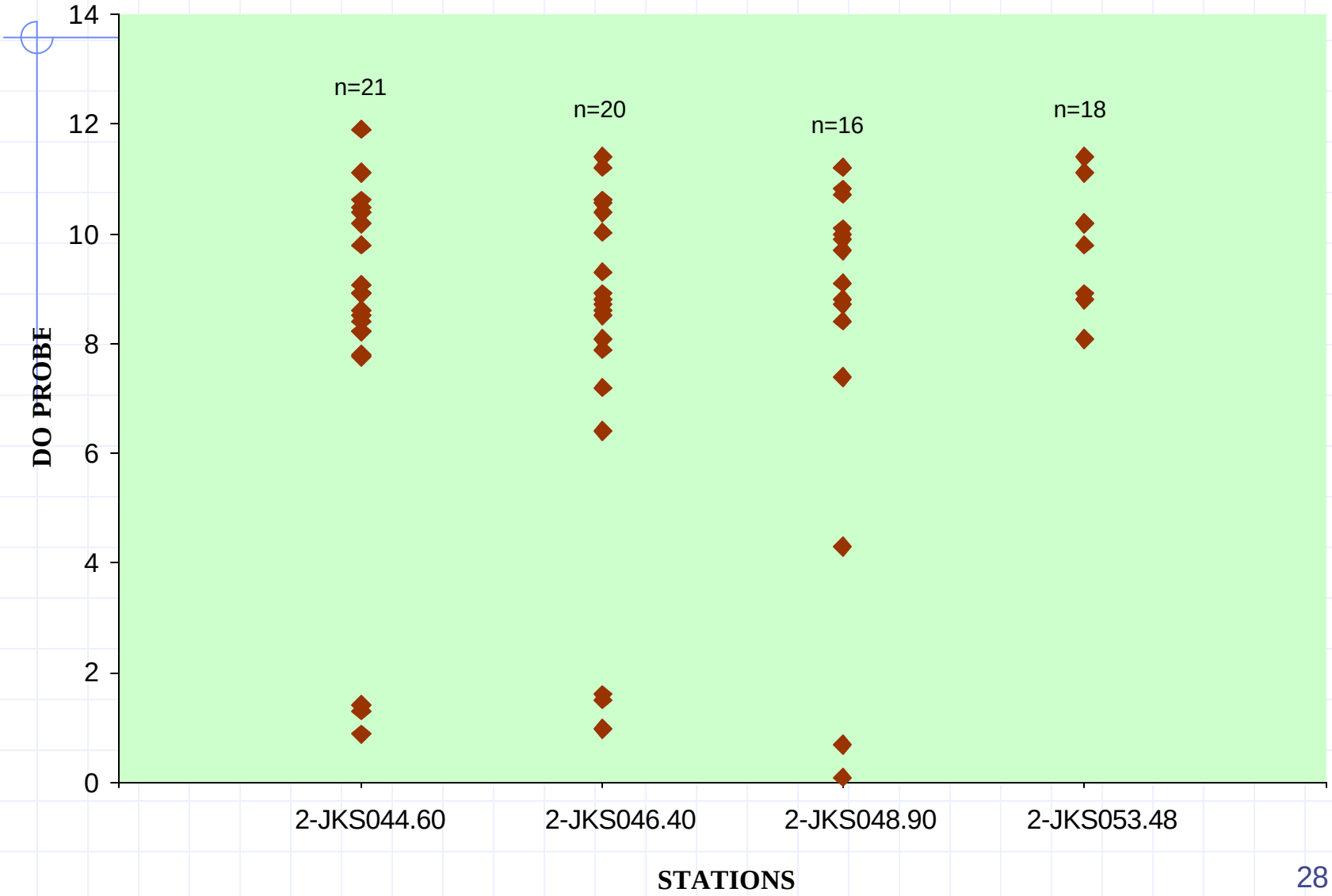
DO data collected at four stations of PHILPOTT RESERVOIR (years 2000, 2001 & 2002)



Evaluate site impairment based on DO data collected at four stations of PHILPOTT RESERVOIR (years 2000, 2001 & 2002)

Station of interest	4ASRE046.90	4ASRE048.98	4ASRE052.31	4ASRE056.06
n	28	31	31	32
No. obs <5	3	5	8	9
sample mean	7.55	7.10	6.66	6.67
sample variance	5.81	8.28	9.56	16.15
Percent exceed the EPA standard	0.11	0.16	0.26	0.28
P-value of Binomial test (H_0: p=0.1 H_a: p>0.1)	0.5406	0.1932	0.0096	0.0033
Bayesian test. (H_0: L=5 H_a: L<5), L is the lower 10th percentile of DO				
With Reference Prior:				
$P(H_0)$	**0.1640**	0.0038	0	0
posterior s.d. of ?	0.6514	0.7325	0.7875	1.008
With Power Prior:				
$P(H_0)$	**0**	0	0	0
posterior s.d. of ?	0.5485	0.5371	0.5439	0.6162

DO data collected at four stations of MOOMAW RESERVOIR (years 2000 & 2001)



Evaluate site impairment based on DO data collected at four stations of MOOMAW RESERVOIR (years 2000 & 2001)

Station of interest	2-JKS044.60	2-JKS046.40	2-JKS048.90	2-JKS053.48
n	21	20	16	8
No. obs <5	3	3	3	0
sample mean	8.16	8.06	8.19	9.81
sample variance	9.73	10.07	12.14	1.32
Percent exceed the EPA standard	0.14	0.15	0.19	0.00
P-value of Binomial test (H_0: p=0.1 H_a: p>0.1)	0.3516	0.3231	0.2108	1.0000
Bayesian test. (H_0: L=5 H_a: L<5), L is the lower 10th percentile of DO				
With Reference Prior:				
$P(H_0)$	0.1497	0.1149	0.1022	0.9968
posterior s.d. of ?	1.0030	1.0500	1.3110	0.7219
With Power Prior:				
$P(H_0)$	0.1338	0.1206	0.1163	0.3301
posterior s.d. of ?	0.6698	0.6832	0.7132	0.7469

Comments

- Advantages
  - Greater flexibility in modeling
  - Allows for site history to be included
  - Can include spatial and temporal components
  - Can better connect to TMDL analysis and probabilistic sampling
- Disadvantage
  - Requires more commitment to the modeling process
  - Greater emphasis on the distributional assumptions